CLAIMS

1. A computer-implemented method for retrieving documents comprising:

inputting the text of one or more documents, wherein each document includes human readable words;

creating context windows around each said word in each document;

generating a statistical evaluation of the characteristics of all of the windows, wherein the results are not a function of the order of the appearance of words within each window; and

combining the results of the statistical evaluation for each window.

2. The method according to Claim 1 further comprising:

determining the likelihood of documents having predetermined characteristics based on the combined statistical evaluation for each window.

3. The method according to Claim 2 further comprising:

assigning a document identifier to each document and context window position; and

determining the document identifier of at least one document having said predetermined characteristics.

4. The method according to Claim 1 further comprising:

defining a plurality of document categories; and

determining the category of a particular document based on the combined statistical evaluation for each window.

5. The method according to Claim 1 further comprising:

determining the word that is in the center of a particular window based on the combined statistical evaluation for each window.

6. The method according to Claim 1 wherein the step of generating a statistical evaluation further includes counting the occurrences of particular words and particular documents and tabulating totals of the counts.

7. The method according to Claim 6 wherein the step of generating a statistical evaluation further includes the step of generating counts about singular word occurrences and about pair-wise occurrences.

8. The method according to Claim 7 further comprising the step of pruning the number of pair-wise counts.

9. The method according to Claim 8 wherein the step of pruning further includes the steps of monitoring the amount of memory used for the pair-wise counts and pruning when a predetermined threshold of memory has been exceeded for the pair-wise counts.

10. The method according to Claim 6 wherein the step of generating a statistical evaluation further includes the step of determining probabilities of particular words appearing in particular documents based on the counts.

11. The method according to Claim 10 wherein the step of generating a statistical evaluation further includes determining conditional probabilities of particular words appearing in particular documents based on the counts.

12. The method according to Claim 11 further comprising the step of calculating a conditional probability based on a Simple Bayes statistical model.

13. The method according to Claim 1 wherein the step of creating context windows around each word further comprises the step of selecting the words appearing before and after each word by a predetermined amount in the document and including those selected words in the window.

14. The method according to Claim 13 wherein the word around which each window is created is not included in the window.

15. The method according to Claim 1 further comprising normalizing the combined results of the statistical evaluation for the windows.

16. The method according to Claim 1 wherein the step of evaluating further comprises, determining a measure of mutual information.

17. The method according to Claim 1 wherein the step of combining includes averaging probability assessments.

18. A computer system comprising:

storage unit for receiving and storing a plurality of documents, wherein each document includes human readable words; means for creating context windows around each said word in each document;

means for generating a statistical evaluation of the content of each window, wherein the order of the appearance of words within each window is not used in the statistical evaluation;

means for combining the results of the statistical evaluation for each window; and

means for determining the probabilities of documents having predetermined characteristics based on the combined statistical evaluation for each window.

19. The computer system according to Claim 18 further comprising:

a document identifier assigned to each document; and

means for determining the document identifier of at least one document having said predetermined characteristics.

20. The computer system according to Claim 18 further comprising:

a plurality of document categories; and

means for determining the category of a particular document based on the combined statistical evaluation for each window.

21. The computer system according to Claim 18 further comprising:

means for determining the word that is in the center of a particular window based on the combined statistical evaluation for each window.

22. The computer system according to Claim 18 wherein the step of generating a statistical evaluation further includes counting the occurrences of particular words and particular documents and tabulating totals of the counts.

23. The computer system according to Claim 22 wherein the means for generating a statistical evaluation further includes means for determining probabilities of particular words appearing in particular documents based on the counts.

24. The computer system according to Claim 23 wherein the means for generating a statistical evaluation further includes means for determining conditional probabilities of particular words appearing in particular documents based on the counts.

25. The computer system according to Claim 18 wherein the means for creating context windows around each word further comprises means for selecting the words appearing before and after each word by a predetermined amount in the document and including those selected words in the window.

26. A computer program product comprising:

a computer program storage device;

computer-readable instructions on the storage device for causing a computer to undertake method acts to facilitate retrieving documents, the method acts comprising:

inputting the text of one or more documents, wherein each document includes human readable words;

creating context windows around each said word in each document;

generating a statistical evaluation of the characteristics of each window, wherein the results are not a function of the order of the appearance of words within each window; and

combining the results of the statistical evaluation for each window.

27. The computer program product according to Claim 26 further comprising:

determining the likelihood of documents having predetermined characteristics based on the combined statistical evaluation for each window.